Web Data Management powered by
Rules and Reasoning in the Semantic Web:
Intelligent Solutions for
Web Data Extraction, Management & Accessibility

Prof. Dr. N. Henze¹
Prof. Dr. G. Gottlob²

¹ University of Hannover, Germany
henze@l3s.de

² Vienna University of Technology, Austria
gottlob@dbai.tuwien.ac.at
Imagine....

- You are a researcher, working for a research institution / university
- You are involved in several projects
- One of the scientific evaluation metrics are your publications
- You put all your publications as your „business card“ on the Web
- You submit all publications developed in the context of a project to the project office
- Somebody at the project offices collects all the publications, and publishes them on the Web
- Similar: each project needs data about you, your affiliation, research background, experience, homepage etc.
The way it is ...

• The information is already on the Web
• We collect, distribute, re–collect, re–distribute information over and over again
  • update, maintenance –> annoying, error–prone

.... the way it should be

• publish information at the – for you – most convenient place, and publish it once!
• let machines crawl, collect and syndicate the data for you, let intelligent user interfaces provide personalized views on the data you are looking for!
Outline

Vision: A Semantic Web

Web content syndication:
- The Personal Reader Framework: a framework for designing, implementing and maintaining Web Content Readers
- A Personal Publication Reader for REWERSE

Web content extraction:
- Automatic extraction of data from Web Sites: Lixto

Summary
What's the main problem in our scenario?

- Machines can't distinguish personal information from publication information – Machines can not understand the *semantics* of the information

- **Vision of a „Semantic Web“**

  "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

The Semantic Web Protocol Stack

by Tim Berners Lee

Diagram showing layers of the Semantic Web Protocol Stack:

1. Unicode
2. URI
3. XML + NS + xmschema
4. RDF + rdfschema
5. Ontology vocabulary
6. Logic
7. Proof
8. Trust
9. Digital Signature
10. Data
11. Rules
– Reasoning on the Web!

• Europe's **top research institutions** and companies with expertise in
  • logic programming, logical and rule-based language
  • reasoning & rules
  • web systems

• joined in a **Network of Excellence**, 6th European Framework Programme

• The **objective of REWERSE** is to establish Europe as a leader in reasoning languages for the Web by
  1. networking and structuring a scientific community that needs it, and by
  2. providing tangible technological bases that do not exist today for an industrial software development of advanced Web systems and applications.
Web Content Syndication:

A Personal Publication Reader
The Personal Reader Framework

• A web content Reader:
  • display web information in a browser

• A Personal Reader:
  • provide a context to each information resource
    • further, detailed information
    • context, background information
    • recommendations for similar information
    • recommendations for potentially interesting additions
    • individually optimized recommendations according to a user's preferences

• Several Personalization Services are available for recommending and providing additional information about information resources on the Web
Content Syndication: Personalization Rules

User Interface & Personalization Services: Personal Publication Reader
Towards generic query, update, and event languages for the Semantic Web.

LOCATION: University of Munich

YEAR: 2004

AUTHORS: Wolfgang May, Jose Julio Alferes, Francois Bry

WORK GROUPS:
- PRA - Presentation, Reviewing and Assessment
- WG ET - University Education and Training
- WG A1 - Web-based Decision Support for Event, Temporal, and Geographical Data
- WG A3 - Personalised Information Systems
- WG E1 - Composition and Typing
- WG E5 - Evolution and Reactivity
- PRA - Presentation, Reviewing and Assessment

WORKING GROUPS:
- Ludwig-Maximilians-Universität München
- Universität Göttingen - Institut für Informatik - Arbeitsgruppe Datenbanken und Informationssysteme

OTHER PUBLICATIONS OF WORKGROUP:
- An Evaluation of Regular Path Expressions with Qualifiers against XML Streams
- Collaborative Categorization on the Web: Approach, Prototype, and Experience Report

OTHER PUBLICATIONS OF AUTHOR:
- A Molecular Biology Database Digest
- A Compositional Semantics for Logic Programs and Reductive Databases
- An Efficient Single-Pass Query Evaluator for XML Data Streams
Towards Generic Query, Update, and Event Languages for the Semantic Web

Wolfgang May, José Júlio Alferes, and François Bry

Institut für Informatik, Universität Göttingen, Germany

CENTRIA, Universidade Nova de Lisboa, Portugal

Institut für Informatik, Ludwig-Maximilians-Universität München, Germany

Abstract. We outline aspects of querying and updating resources on the Web and on the Semantic Web, including the development of query and update languages in course of the REWERSE project. When considering updates and communication of updates between autonomous sources, reactive behavior plays an important role such that an event language is required. This article provides a systematic outline of the intended research steps towards handling reactivity and evolution on the Web.

1 Introduction

Use of the Web today—commonly known as the “World Wide Web”—mostly focuses on the page-oriented perspective: most of the Web consists of browsable HTML pages only. From this point of view, the Web can be seen as a graph that consists of the resources as nodes, and the hyperlinks form the edges. Here, queries are stated against individual nodes, or against several nodes, e.g., with formalisms like F-Logic [16] or Lixto [3]; or in case that the sources are provided in XHTML, they can be queried by XQuery, XPathLog [17], or Xcerpt [5]. As such, the Web is mainly seen from its static perspective of autonomous sources, whereas the behavior of the sources, including active roles of resources plays not any important role here.

But there is more on the Web of today than HTML pages. Leaving the superficial point of view of HTML pages, the Web can be seen as a set of data...
Towards generic query, update, and event languages for the Semantic Web.

AUTHORS: Wolfgang May, José Julio Alferes, François Bry

LOCATION: University of Munich

YEAR: 2004

François Bry

Professor
Teaching and Research Unit "Programming and Modelling Languages"
Computer Science Institute

http://www.personal.uni-muenchen.de/8060/pms/
Towards generic query, update, and evolution in the Semantic Web.

Objective: WG A3 - Personalized Information Systems

The working group A4 on "personalized information systems" aims at advancing the state of the art of customized information delivery applications in the Semantic Web. Applications like adaptive information systems and personalized e-commerce systems use backbones built upon reasoning. Results from working groups IL-L5 will bring decisive input for bringing personalization to the Semantic Web and thus advance the design of the Adaptive Web.
Web Content Extraction
Electronic Commerce

E-procurement, quality assurance:

Quality norm 06865 for tires

OEM PORTAL

IVECO

- Pirelli
- Michelin
- Avon
- Dunlop
- Firestone
At what prices are our tires and those of our competitors sold? React to changes...
Marketing & Business Intelligence

Oracle 9

Marketing Department

Business Objects report

BI Tool
The Wall

Problem: Make web contents accessible to electronic data processing

WEB
HTML pages
layout

Corporate edp apps
structured data, Databases,
XML
WEB
HTML pages
layout

Corporate edp apps
structured data, Databases, XML
Web wrapping

**Goal:** Make web contents accessible to electronic data processing

- Web
  - HTML pages
  - Layout
- Corporate edp apps
  - Structured data
  - Databases
  - XML
Web wrapping

Goal: Make web contents accessible to electronic data processing

Wrappers: HTML → select → extract → annotate → XML
### Tyre Offers

<table>
<thead>
<tr>
<th>Brand</th>
<th>Profile</th>
<th>Size</th>
<th>Speed</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goodyear**</td>
<td>EAGLE NCT 5</td>
<td>195/65 R15 91V</td>
<td></td>
<td>£44,10</td>
</tr>
<tr>
<td>Pirelli***</td>
<td>P 6000 Powergy</td>
<td>195/65 R15 91H</td>
<td></td>
<td>£44,70</td>
</tr>
<tr>
<td>Continental***</td>
<td>EcoContact CP</td>
<td>195/65 R15 91H runout</td>
<td></td>
<td>£46,00</td>
</tr>
<tr>
<td>Bridgestone***</td>
<td>Turanza ER 91</td>
<td>195/65 R15 91H</td>
<td></td>
<td>£46,50</td>
</tr>
<tr>
<td>Michelin***</td>
<td>Pilot PRIMACY</td>
<td>195/65 R15 91H</td>
<td></td>
<td>£56,40</td>
</tr>
<tr>
<td>Goodyear***</td>
<td>EAGLE NCT 5</td>
<td>195/65 R15 91H</td>
<td></td>
<td>£44,20</td>
</tr>
<tr>
<td>Pirelli***</td>
<td>P 6000 Powergy</td>
<td>195/65 R15 91H</td>
<td></td>
<td>£46,70</td>
</tr>
<tr>
<td>Dunlop***</td>
<td>SP SPORT 200</td>
<td>195/65 R15 91H E</td>
<td></td>
<td>£46,40</td>
</tr>
<tr>
<td>Pirelli***</td>
<td>P 6</td>
<td>195/65 R15 91H</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10 tyres from 118 were displayed. (1 - 10)

---

Prices includes postage, packing and VAT within mainland UK.

***Please note: these tyres are subject to a delivery period of up to***
Patterns:

<table>
<thead>
<tr>
<th>Tyre</th>
<th>Brand</th>
<th>Profile</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bridgestone**</td>
<td></td>
<td>Turanza ER 31</td>
<td>£46.50</td>
</tr>
<tr>
<td>Goodyear**</td>
<td>EAGLE</td>
<td>NCT 5</td>
<td>£44.20</td>
</tr>
<tr>
<td>Pirelli**</td>
<td>P 6000 Powergy</td>
<td></td>
<td>£44.70</td>
</tr>
<tr>
<td>Dunlop**</td>
<td>SP SPORT</td>
<td>01</td>
<td>£44.30</td>
</tr>
</tbody>
</table>

10 tyres from 118 were displayed. (1 - 10)

Prices includes postage, packing and VAT within mainland UK.

***Please note: these tyres are subject to a delivery period of up to 1 week.
<?xml version="1.0" encoding="UTF-8"?>
<document>
  <record>
    <number>409449118</number>
    <item>98 Degrees - Notebook - New</item>
    <picture/>
    <price>2.99</price>
    <currency>$</currency>
    <bids></bids>
  </record>
  <record>
    <number>413171469</number>
    <item>Notebook - Compaq Presario 1207</item>
    <price>730.00</price>
    <currency>AU $</currency>
  </record>
  [...]
</document>
<table>
<thead>
<tr>
<th>Item Description</th>
<th>Price</th>
<th># of Bids</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEC LAPTOP/NOTEBOOK PII-233.96MB, 13&quot;, CD</td>
<td>$399.00</td>
<td>7</td>
<td>Jul-13 10:14</td>
</tr>
<tr>
<td>NEC Versa S/5a 486 DX 12 MB RAM 340 MB HD</td>
<td>$20.60</td>
<td></td>
<td>Jul-17 09:57</td>
</tr>
<tr>
<td>NEC READY 120LT NOTEBOOK COMPUTER</td>
<td>$350.00</td>
<td></td>
<td>Jul-17 09:44</td>
</tr>
<tr>
<td>NEC 200C W/IBM ETHERJET CARD1</td>
<td>$50.00</td>
<td></td>
<td>Jul-17 09:30</td>
</tr>
<tr>
<td>IBM 760XL CD, FLOPPY AND MORE!</td>
<td>$99.00</td>
<td></td>
<td>Jul-17 09:12</td>
</tr>
<tr>
<td>Qty 10 NEC Versa LX PII 300Mhz/128M warranty</td>
<td>$4400.00</td>
<td></td>
<td>Jul-17 09:04</td>
</tr>
</tbody>
</table>

For more items in this category, click these pages:
1 = 2 3 4 5 (next page)
Different Approaches so far:

- **Programming** (Java, Perl, WebL, SQL+...)
  - very complicated & boring & expensive
  - testing very difficult
- **Screen scrapers**
  - no complex data structures extracted
  - nontrivial extraction task require complicated procedural programming
  - wrappers not robust w.r.t. change
- **Wrapper induction**
  - requires larger amounts of sample data
  - accuracy not satisfactory in all situations
  - current systems text-based (not tree-based)
Extraction vs. Querying and Transforming

**HTML** → **XML** → **XSLT** → **RDF**

**DATA EXTRACTION**
- selection
- annotation
- hierarchical (re)grouping

**DATA TRANSFORMATION**
- combination (joins, etc.)
- querying (XSLT, XQUERY)
- personalization
Extraction vs. Querying and Transforming

DATA EXTRACTION
- selection
- annotation
- hierarchical (re)grouping

DATA TRANSFORMATION
- combination (joins, etc.)
- querying (XSLT, XQUERY)
- personalization

HTML ➔ XML ➔ XSLT ➔ RDF

Visual Wrapper
Transformation Server
SHORT DEMO

www.paris.org...
Thank you very much for your attention!

Visit us in the EU Village, Stand D

www.lixto.com
www.personal-reader.de